Reprinted from the PROCEEDINGS OF THE NATIONAL ACADEMY OF SCIENCES
Vol. 53, No. 1, pp. 134–139. January, 1965.

*TOPOLOGICAL MAPPING OF ORGANIC MOLECULES**

BY JOSHUA LEDERBERG

DEPARTMENT OF GENETICS, STANFORD UNIVERSITY SCHOOL OF MEDICINE, PALO ALTO

Communicated November 23, 1964

A structural formula in organic chemistry is a statement of the topological connectivity of the atoms of a compound. However, the topological theory of organic chemical structure has not been formally developed. Partly in consequence, the taxonomy, i.e., nomenclature, notation, and homology, of the field lags behind its substance which impedes communication, whether this be information retrieval or professional education. Witness the mystification often provoked by the proper name of a new drug. The need for a more complete formal system became acutely evident in an effort to write computer programs for the logical analysis of mass spectra.[1, 2] It was found that the mapping of organic structures on standardized forms contributed to the simplification of the problem and this will be illustrated here.

Tree Structures.—Acyclic molecules are easy to standardize, but topological principles are hardly used in current practice. Over thirty years ago, Henze and Blair[3] pointed out, in their enumeration of alkanes, that a unique centroid can be found in any chemical tree. This is either a link that evenly divides the skeleton of the tree, or a single atom each branch from which carries less than half of the skeletal atoms. The unique centroid is then the starting point for a canonical mapping of the tree, following simple rules of precedence of the constituent radicals according to their composition and topological structure. A compact, unique, and unambiguous notational system[2] has been established from these canons and need detain us no further here.

Cyclic Structures.—Rings are much more difficult to process on a node-by-node basis. Ambiguities due to symmetry are usual, and many paths can be evaluated

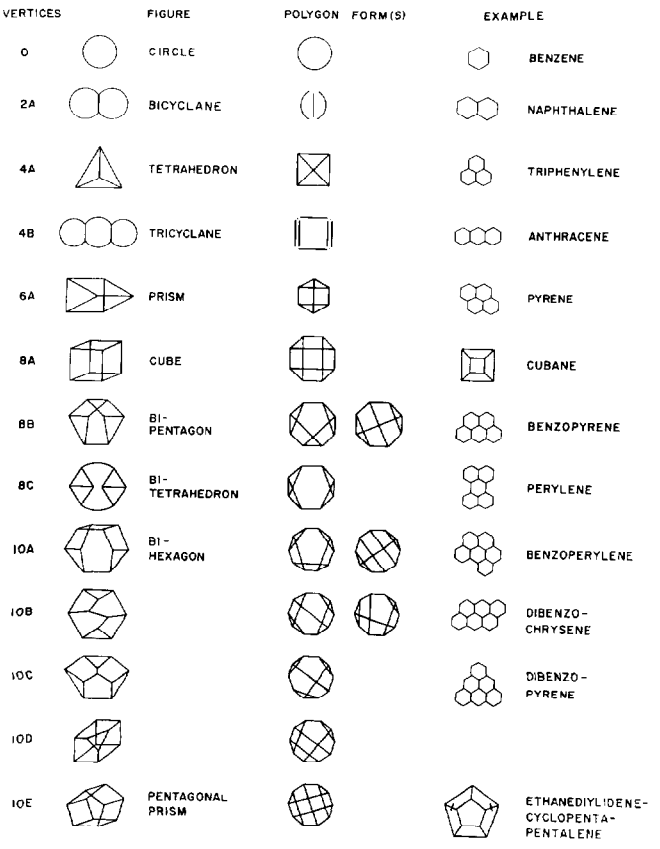


FIG. 1.—Fundamental trivalent polyhedra, including degenerate forms, with up to 10 vertices. Examples are drawn from polycyclic compounds where possible; these do show an unusual degree of symmetry.

only by recursively searching through the entire graph. This approach was therefore abandoned in favor of a fundamental classification of the graphs. To achieve this, a number of simplifying steps are introduced. The *first* of these is to isolate the paths within the ring. The classification then depends on the set of branch points. Organic rings rarely have more than three branches at any point; an instance of four branches can be accommodated by exception. A *second* simplification then asks only for a classification of regular trivalent graphs. How, then, can the set of trivalent graphs be systematically arranged, and how can we be assured of having deduced the entire set, without isomorphic redundancies? The graphs of Figure 2 constitute such a set, of order 8, except for the gauche forms discussed later. Similar sets of orders 10 and 12 have also been generated on the computer.

Polygonal graphs are relatively easy to compute, but they fail to show many of the symmetries of the figures. This is dramatized by the two isomorphic polygonal representations of the bipentagon. Furthermore, not all the graphs have Hamilton circuits (i.e., can be represented as chorded polygons). Some, like 8 M, require additional vertices, and these are not so readily generated by a polygon algorithm.

A *third* basis was therefore introduced, the trivalent graphs being identified with *polyhedra,* including some degenerate forms and derivations from them. As shown

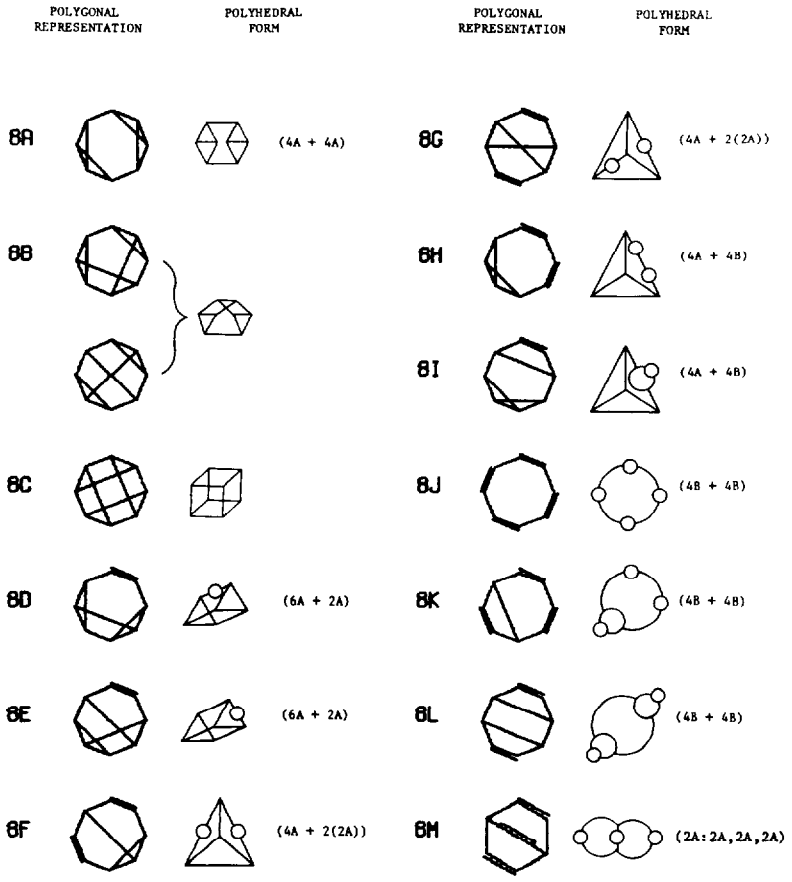
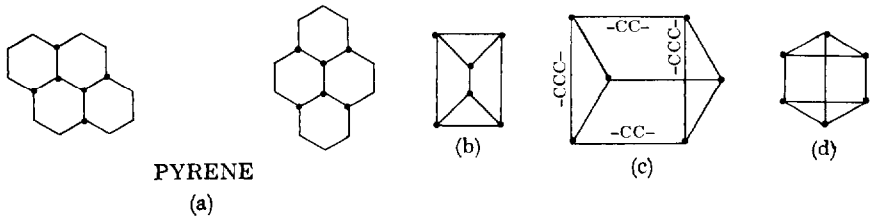


FIG. 2.—Trivalent graphs of order 8. *8A–C* are fundamental forms already seen in Fig. 1. Polygonal representations are computer-generated plots; corresponding polyhedral presentations were drawn by hand. Adjacent to each of the unions is a code for its composition.

in Figure 1, the formulation of polyhedra emphasizes the orderly development of the set of graphs and the symmetries of each structure, and thus facilitates the recognition of isomorphisms.

Polyhedral Forms.—For topological analysis of a ring the linear paths and the vertices connecting them are first identified. The vertices are simply the branch points, i.e., the atoms with three or more links to the rest of the ensemble. For these purposes a double or triple bond is a single link. The paths are then the intervals between the vertices. A path may be a simple link or a linear string of tandemly



linked atoms. For example, marking the paths of pyrene, (*a*), gives the diagram (*b*) which is readily recognized as isomorphic to the prism (*c*) and its formal graph (*d*). The isomorphism of (*b*) with (*c*) could also be established algorithmically by systematic permutation of the incidence matrix of the graphs.

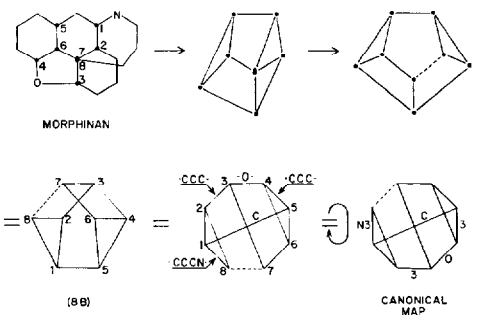


FIG. 3.—Mapping a complex ring; morphinan.

Figure 1 lists convex trivalent polyhedra with up to 10 vertices, on which structures with up to six rings can be mapped. It also includes some laminar forms, e.g., the circle, bicyclane, and tricyclane, which might be regarded as degenerate polyhedra with 1, 2, or 3 faces, respectively. The series has also been algorithmically expanded further in a computer program.

Any trivalent graph is assumed to represent either a polyhedron[5] or a gauche graph of the same order, or the union of two or more graphs of lower order. A union is obtained across a pair of cut edges of two graphs. The derived forms are classified according to the largest polyhedron or symmetrical union, e.g., bitetrahedron, contained in the graph.

Spiro-Atoms.—A quadrivalent vertex is mapped as a collapsed *edge* of a trivalent graph: $>\cdot-\cdot< = >\cdot<$. The parent graph is almost always subject to two or three choices. That partition is chosen which leads to the least complex map, i.e., the nearest to a polyhedron with the least appendages. Figure 3 illustrates a mapping of morphinan.

Gauche Graphs.—The Ring Index[4] with its 11,524 examples of rings known to organic chemistry contains no example of a finite gauche graph, i.e., one whose representation on the plane has obligatory crossed paths. (Optional crossed-path formulas are sometimes preferred to show the homology of figures to one another.) A theorem of Kuratowski has shown that a gauche graph must contain either Figure 4*a* or 4*b*;[6] Figure 4*a* may be discounted as an unlikely pentaspiro complex. Is the nonexistence of Figure 4*b* a coincidence? Its representation, Figure 4*c*, as an internally chorded tetrahedron may throw some light on this. A gauche structure would require access of a chemical path to the interior of a urotropinelike molecule, Figure 4*d*. However, with the interposition of longer paths, it should be possible to fill this topologicochemical hiatus.

Limitations of Topological Description.—Mapping is intended to convey only the connectivity relationship of a set of atoms, which is only the first-order account of a molecular structure. Information on the bond character or stereochemical ordering of links must be furnished in addition. At least the latter is not difficult: the main programming problems involve accounting for all the symmetries. Molecular

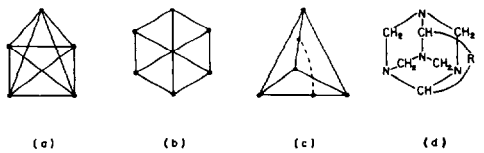


FIG. 4.—Gauche graphs: (*a*) and (*b*), Kuratowski's fundamental forms of nonplanar graphs; (*c*), a three-dimensional representation of (*b*); and (*d*) is a hypothetical example.

conformations represent another domain, whose computation belongs mainly to numerical analysis rather than topology. Thus, from the standpoint of the present paper, interlocked rings have the same connectivity as separate rings. The catena structures cannot, however, be properly drawn without crossed paths. Many real conformations may also have to be shown with superimposed paths in any actual projection on the plane.

Applications of Topological Mapping.—The primary purpose of this analysis was to provide a framework for computable logic in organic chemistry, especially the analysis of mass spectra. The theoretical ideas of this presentation are very primitive and its main virtue may be to provoke a more sophisticated mathematical formulation.

Once a standard form is chosen for mapping, canons can be elaborated for the ordering of paths leading in turn to a systematic, compact, computable notation for organic structures.[2] The main burden of the standardization is a dissection of the symmetries of the diagram, then a rule of choice among the permutations of the labels. Thus, in the example of Figure 3, the diagram 8B has fourfold symmetry. The symmetry permutations of the principal polygon can be expressed, with the corresponding path lists, as:

Path	[12345678	(28) (37) (46)	(15) (26) (39) (48)	(15) (24) (38)
12	—	NCCC	—	CCC
23	CCC	Fused	—	O
34	O	—	Fused	CCC
45	CCC	—	CCCN	—
56	—	CCC	—	NCCC
67	—	O	CCC	Fused
78	Fused	CCC	O	—
81	CCCN	—	CCC	—
15	C	C	C	C
28	—	—	—	—
37	—	—	—	—
46	—	—	—	—

The top heavy path list of (28) (37) (46) makes this the canonical choice. A linear code for the mapping of morphinan is then (8H NCCC,$,—,—,CCC,O,CCC,—,C,—, —,—), or morecompa ctly, (8H N3,$,,,3,0,3,,1), the map being thus reduced to a twelve-dimensional vector. A computer program would unambiguously recognize any of the permuted path lists as equivalent forms, and can perform the tedious exercise just concluded.

More important than the notation, this framework enables a computer program to generate hypotheses of organic molecular structure in an algorithmic, exhaustive, and, above all, redundancy-free sequence, important if the computer is to amplify human logical capacity in this field.

Summary.—An algorithmic approach to the topological mapping of organic molecules is presented. Three structures are initiated at a unique centroid of the skeletal atoms. Cyclic structures are more difficult. However, the set of regular graphs of degree 3 can be generated on a basic set of polyhedra. Any organic ring molecule can be mapped on one of these graphs. Exceptional quadrivalent vertices (spiro fusions) are expanded to a pair of trivalent vertices. Gauche graphs, with obligatory crossed paths, have not yet been realized in organic chemistry, probably owing to the difficulty of accessing a chemical path as an interior chord of a closed molecule.

* Research connected with this program has been supported by grants from the NIH (NB-04270 and AI-05160), National Science Foundation (G-6411), and NASA (NsG 81-60).

[1] Lederberg, J., *Computation of Molecular Formulas for Mass Spectrometry* (San Francisco: Holden-Day, 1964).

[2] Lederberg, J., *DENDRAL-64, A System for Computer Construction, Enumeration and Notation of Organic Molecules as Tree Structures* (NASA CR 57029).

[3] Henze, H. R., and C. M. Blair, *J. Am. Chem. Soc.*, **53**, 3077 (1931).

[4] *The Ring Index* (American Chemical Society, 1964), 2nd ed., suppl. 2.

[5] All the polyhedra (but only some of the derived forms) have a polygonal representation, i.e., a Hamilton circuit, which greatly facilitates the computation of these graphs, as it does in many applications of graph theory. Tait[7] had conjectured that any true, trivalent polyhedron had a Hamilton circuit. Tutte[8] has demonstrated a counterexample with 46 vertices. If this is the smallest exception, the conjecture is an adequate basis for present purposes. No examples of chemical rings actually recorded in the *Ring Index* defy descriptions in this system.

[6] Berge, C., *The Theory of Graphs* (New York: John Wiley & Sons, 1958), chap. 21.

[7] Tait, P. G., *Phil. Mag.* (Series 5), **17**, 30 (1884).

[8] Tutte, W. T., *J. London Math. Soc.*, **21**, 98 (1946).